Sentiment Analysis of Lok Sabha Election

B.Tech Computer Science Engineering, 7th Semester, 2022-23

Mentor

Prof. Sudipta Bhattacharya
Assistant Professor
Dept. of CSE & IT
Bengal Institute of Technology, Kolkata

Group Members

Name	MAKAUT Roll No.
Jony Paul	12100119073
Imon Bera	12100119075
Shouvik Sarkar	12100119088
Dibyendu Nandan	12100119110
Bitan Ranjan Maiti	12100119115

BENGAL INSTITUTE OF TECHNOLOGY
1 GOVT. COLONY, TECH TOWN, DHAPA MANIPUR, ON BASANTI HIGHWAY, HADIA, KOLKATA-150

Acknowledgement

This project would not been a successful without the sincere cooperation and guidance of our mentor Prof. Sudipta Bhattacharya, Assistant Professor, Dept of CSE & IT, Bengal Institute of Technology, Kolkata, who has provided us with useful resources and motivation through the various phases of this project which has proven to be crucial for the completion of this documentation.

Signature of Group Members:

Name: Jony Paul
MAKAUT Roll No. 12100119073

Name: Imon Bera
MAKAUT Roll No. 12100119075

Name: Shouvik Sarkar
MAKAUT Roll No. 12100119088

Name: Dibyendu Nandan
MAKAUT Roll No. 12100119110

Name: Bitan Ranjan Maiti
MAKAUT Roll No. 12100119115

Prof. Sudipta Bhattacharya
Assistant Professor
Dept. of CSE & IT
Bengal Institute of Technology

Signature of HOD

Dr. Shanta Phani
Head of the Department
Dept. of CSE & IT
Bengal Institute of Technology

TABLE OF CONTENTS

Abstract:

Social media, a gift of the internet, has the ability to create a wide range of impact on people through sharing and voicing individual opinions. Huge sets of data, generated by various social media platforms, can be used in miscellaneous data analysis techniques if dealt properly. Such a process is Sentiment Analysis , through which we have tried to understand the diverse political trend of digital India. We have focused on two separate trends within our national politics (BJP and INC). We have gathered target tweets from Kaggle and used GloVe on it to extract accurate features. We have built an LSTM model to identify positive, neural and negative sentiments of the tweets and used the model to predict sentiments on recent Twitter data in order to determine which political party is currently performing best for the public.

Introduction:

Sentiment Analysis is the process of collecting and analyzing data based upon the person's feelings, reviews and thoughts. Sentimental analysis is often called opinion mining as it mines the important features from people's opinions. Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of features from a large data set.

Sentiment Analysis has various applications. It is used to generate opinions for people on social media by analyzing their feelings or thoughts which they provide in the form of text. Sentimental Analysis is used in many real life scenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or marketing.

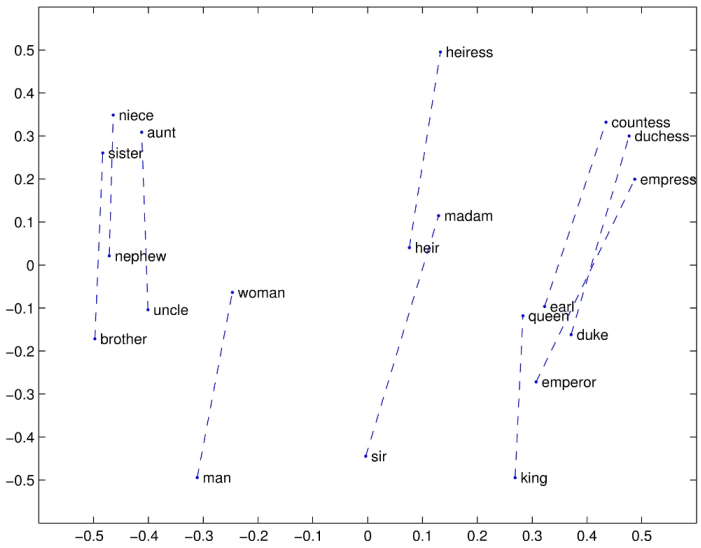
Natural Language Toolkit (NLTK) is a library in Python, which provides a base for building programs and classification of data. NLTK is a collection of resources for Python that can be used for text processing, classification, tagging and tokenization. This toolbox plays a key role in transforming the text data in the tweets into a format that can be used to extract sentiment from them. NLTK provides various functions which are used in pre-processing of data so that data available from twitter become fit for mining and extracting features.

OneHotEncoder is a categorical feature encoding technique in machine learning. It transforms categorical variables into a binary vector representation, where each category is represented by a binary value (0 or 1). Each category becomes a separate feature, eliminating the issue of ordinality in categorical variables. OneHotEncoder is commonly used as a preprocessing step to convert categorical data into a format suitable for machine learning algorithms that require numerical inputs.

In NLP models, we deal with texts which are human-readable and understandable. But the machine doesn't understand texts, it only understands numbers. Thus, word embedding is the technique to convert each word into an equivalent float vector.

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

GloVe is designed in order that such vector differences capture as much as possible the meaning specified by the position of two words. Such one example is below —



The underlying concept that distinguishes man from woman, i.e. sex or gender, may be equivalently specified by various other word pairs, such as king and queen or brother and sister. To state this observation mathematically, we might expect that the vector differences between man - woman, king - queen, and brother - sister might all be roughly equal. This property and other interesting patterns can be observed in the above set of visualizations.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a pre-trained sentiment analysis model specifically designed for analyzing sentiment in social media texts. It is a rule-based model that assigns sentiment scores to individual words and combines them to calculate the overall sentiment polarity of a given text. VADER considers the intensity of sentiment, handles negations, punctuation, capitalization, and incorporates sentiment-related features such as emoticons and slang commonly found in social media content. It is widely used due to its effectiveness in capturing sentiment nuances and its ability to handle informal and short texts often encountered in platforms like Twitter or online reviews.

Related Works:

In the past, countless research papers have been published on the subject of sentiment analysis. In this age of the internet where social networking websites like Facebook, Twitter, Instagram are thriving, it is obvious performing sentiment analysis on such huge generated data will be a hot topic. While the intention of such analysis is sometimes field specific, most of the time such analysis is being done to identify how much something specific is in favor or against the population, and the polarity of it. Let's take a look to a few such similar works—

B. Pooja et al [3] are the ones who did Sentiment Analysis as well as Emotional Analysis (eg.., empty , sadness , anger etc..,) upon user opinions from Twitter data. Their aim was to prove such analysis can serve a gateway for consumer needs and generate growth opportunities in businesses. They used various machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM) and various libraries such as SentiWordNet , Matplotlib , TextBlob etc. Based on their comparison among their train data and test data, they reached a conclusion that Naive Bayes outperforms all the other algorithms with an accuracy of 96%.

M. Lamberti et al [4] collected their dataset from mainly two sources, Kaggle and Sentiment140. The initial difficulty they were trying to solve was removal of emoticons, acronyms and sequence of repeated characters since twitter comments are usually infested with such slangs. They easily overcame the problem by using an emoticon dictionary , an acronym dictionary and a stop word dictionary during pre-processing of data. As per machine learning algorithms they used Naive Bayes, Multi-variate Bernoulli, Multinomial and Gaussian models. Later they introduced language models such as N-Gram model, unigram model, bigram model etc to see a slight increase in accuracy of 0.02 score.

D. Manu et al [5] was centered around analyzing Lok Sabha Election 2019, India. Their aim was to find out how conversations around the internet formed during the Lok Sabha Election 2019 and what impacts it had on real life. They started with a very thorough analysis of the history of politics in India and what effects it has on the normal population. For the data analysis purpose they collected a huge amount of 45 million tweets from the 2019 era. They used LDA Topic Modelling, Word Cloud Analysis , manifesto Analysis to shape the data and take a look from within. They used a network analysis tool - Gephi to create a complete visualization of the data. On the basis of their analysis they came to a conclusion that during the 2019 election, BJP's popularity surpassed that of INC's by huge numbers hence being at an advantage against their political rivals.

Proposed Methodology:

The model training procedure involved using a labeled political dataset called "Twitter Sentiment Dataset (2021)" from Kaggle, containing 162,973 tweets with positive, neutral, and negative sentiments. To address the dataset's imbalance, each sentiment category was limited to 35,500 tweets by shuffling the dataset.

For collecting the prediction dataset, initially, tweets were scraped using Tweepy and a Twitter Developer account. However, due to limitations in the Twitter API V1.0, only a small number of tweets could be collected. Later, the raw data was merged with a publicly available Kaggle dataset called "Dataset of Indian Politics Tweets & Reactions," consisting of 50,000 unlabelled tweets from various Indian political parties including BJP and INC.

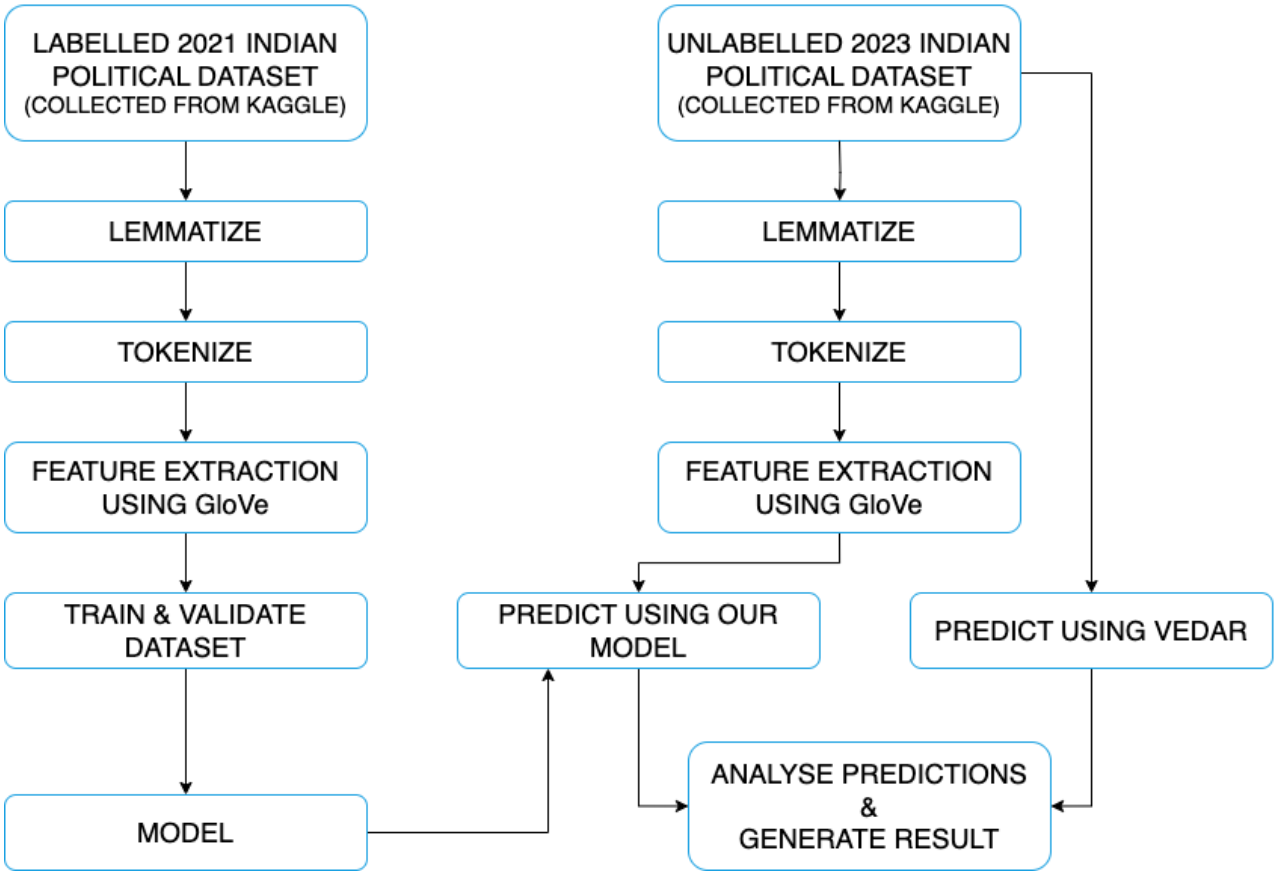
Web scraping was performed using specific keywords for each political party to separate the prediction dataframe. Various NLTK libraries and methods were utilized for preprocessing the training data, including normalization, lemmatization, tokenization, and lowercasing.

GloVe was used to create a word dictionary, enabling the extraction of features from the preprocessed data. OneHotEncoder was applied to the training labels for appropriate fitting. Padding was employed to transform the variable-length tweets into a fixed shape, which is a necessary step in model fitting.

A Sequential LSTM-based model was used to train and fit the data, with the training dataset split to include a validation set for model evaluation. Classification reports and confusion matrices were generated to visualize the model's performance.

Afterwards, the prediction data was separated using keywords related to each major political party (BJP and INC). These respective datasets were preprocessed and passed through the trained model to obtain the prediction dataframe.

To compare the model's accuracy, VADER sentiment analysis was applied to the same prediction dataset, and the results were plotted in separate pie charts for comparison.



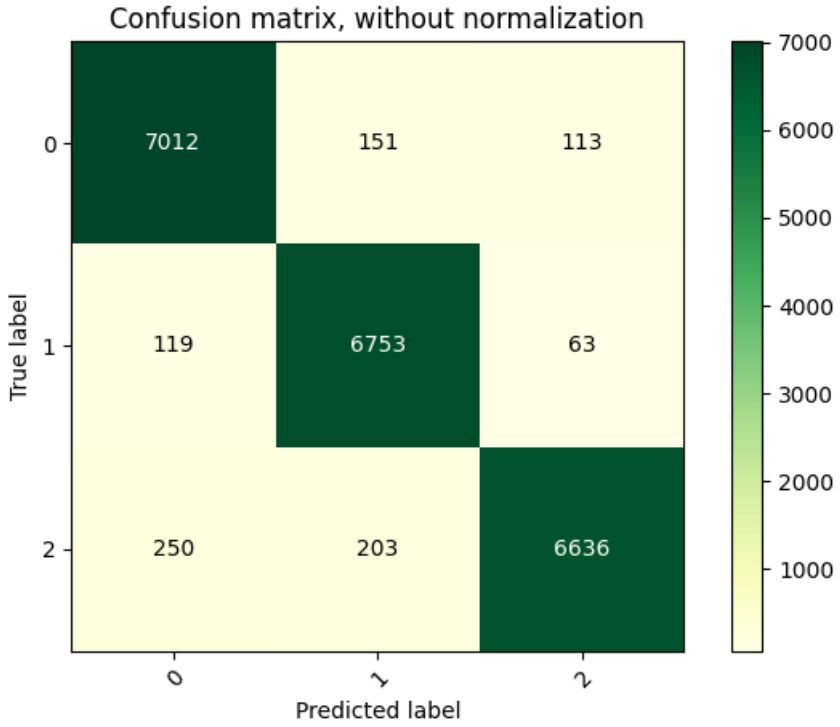
FLOW CHART OF METHODOLOGY

Results and Discussion:

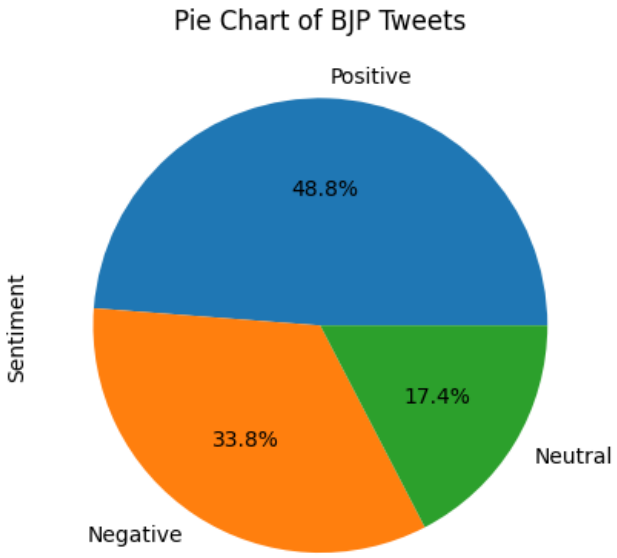
Upon completion of the training of the model on a labeled political dataset and evaluating it using a validation set, we generated a classification report. This report provides insights into the performance of our model in predicting sentiment categories (positive, neutral, negative) for political tweets. The classification report includes metrics such as precision, recall, F1-score, and support for each sentiment category. It helps us understand the accuracy and effectiveness of our model in classifying sentiments.

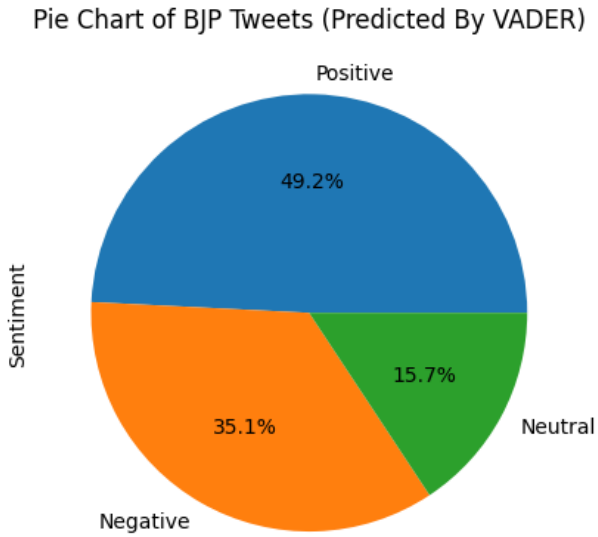
	precision	recall	f1-score	support
0	0.95	0.96	0.96	7276
1	0.95	0.97	0.96	6935
2	0.97	0.94	0.95	7089
accuracy			0.96	21300
macro avg	0.96	0.96	0.96	21300
weighted avg	0.96	0.96	0.96	21300

Additionally, a confusion matrix was constructed to visually represent the performance of our model. The confusion matrix illustrates the number of correctly and incorrectly classified instances for each sentiment category. It provides a deeper understanding of the model's predictive capabilities and highlights any areas where the model may struggle in distinguishing between certain sentiments.

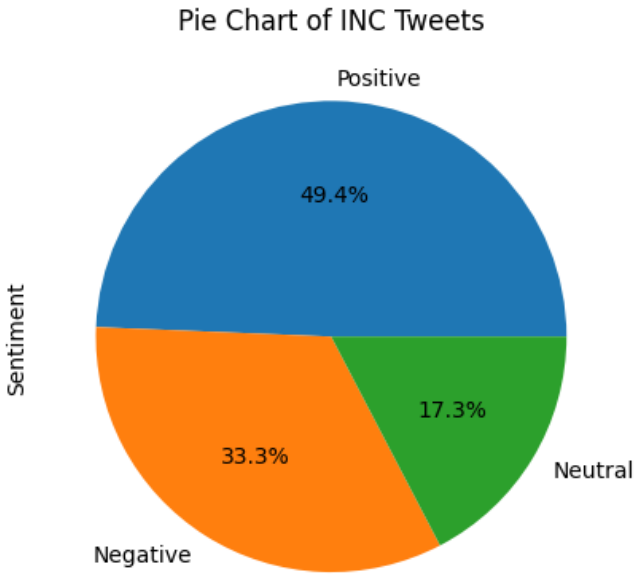


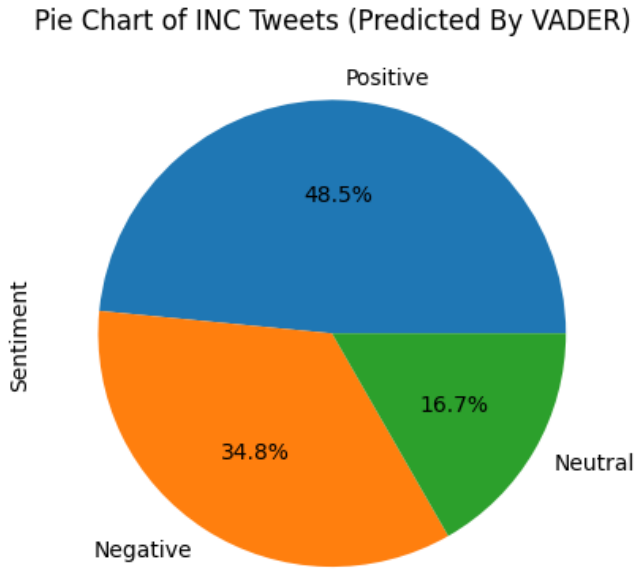
To compare our model's predictions with the VADER sentiment analysis, we created pie charts displaying the distribution of sentiments (positive, neutral, negative) for each political team. The pie charts represent the proportions of sentiment categories predicted by our model and VADER on the same prediction dataset.



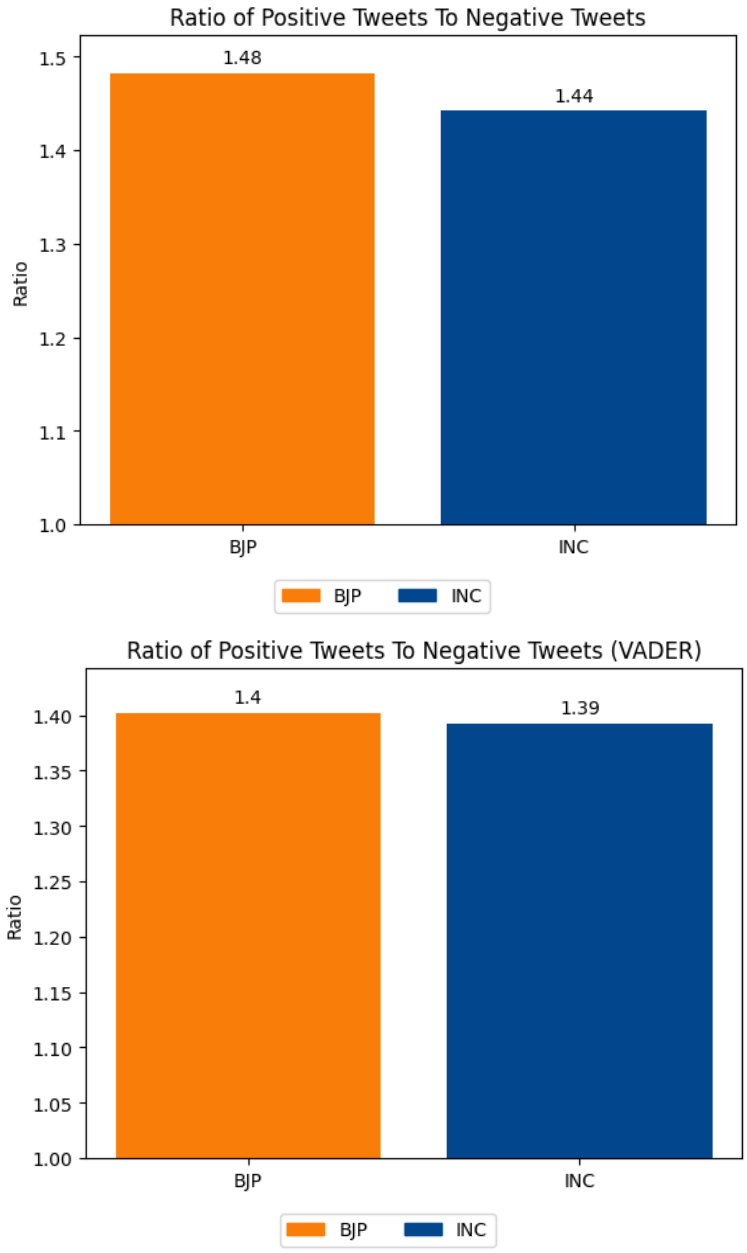


By visually comparing the two pie charts, we can observe the similarities and differences in sentiment distributions. This comparison helps us assess the alignment between our model's predictions and VADER's predictions for the given political teams. It provides insights into the model's performance relative to an established sentiment analysis tool like VADER.





We plotted a bar graph to compare the ratio of positive to negative sentiments between the two political parties using the prediction data from our model. Similarly, we repeated this process using VADER predictions. The bar graph allows us to visualize and compare the sentiment distributions for the political teams.



By examining the bar graph, we can identify any variations in the positive-to-negative sentiment ratio between the two political parties. This analysis provides a comparative view of the sentiment tendencies towards each political team as predicted by our model. Comparing the model's results with VADER's results further enables us to assess the consistency and reliability of our model's sentiment predictions.

Conclusion:

Relevance of Social Media Analysis: The sentiment analysis of Twitter data demonstrates the importance of analyzing social media platforms to gain valuable insights into public opinion. With its widespread usage for political discourse, Twitter serves as a valuable source of data to understand sentiment towards political parties and related topics. By harnessing social media analysis, researchers and stakeholders can access real-time feedback and gauge public sentiment in an efficient manner.

Diversity of Sentiments: The analysis unveils the diverse range of sentiments expressed on social media regarding Indian political parties. Public opinion showcases a mix of support, criticism, and neutral sentiments towards different parties and their policies. This diversity highlights the complexity of public sentiment and reinforces the need for a comprehensive understanding of the various perspectives held by individuals across the political spectrum.

Limitations and Challenges: The project sheds light on the limitations and challenges associated with sentiment analysis on social media. It is crucial to recognize the potential biases inherent to these platforms, such as echo chambers, filter bubbles, and the propensity for polarized discussions. Moreover, while sentiment analysis provides valuable insights, it is important to acknowledge that social media data may not be representative of the entire population's views and should be interpreted cautiously.

Political Context: The sentiment analysis needs to be interpreted within the political context of the time. Political parties and their policies cannot be fully understood without considering the broader political landscape, historical events, and ongoing developments. The dynamics of political alliances, regional issues, and national concerns shape public sentiment and can significantly influence the sentiments expressed on social media platforms.

By considering the relevance of social media analysis, acknowledging the diversity of sentiments, recognizing the limitations and challenges, and contextualizing the analysis within the political landscape, researchers and stakeholders can extract valuable insights to inform decision-making, communication strategies, and policy formulations.

Future Scope:

Some of future scopes that can be included in our research work are:

Enhancing Model Performance: Continuously work on improving the accuracy and performance of the sentiment analysis model by experimenting with different deep learning architectures, such as Bidirectional LSTM, GRU, or Transformer models like BERT or GPT. Fine-tuning these models on domain-specific data can potentially lead to better sentiment classification results.

Multilingual Sentiment Analysis: Extend the project to analyze sentiments in multiple languages, considering the diverse linguistic landscape in India. This could involve training language-specific sentiment analysis models or leveraging pre-trained multilingual models like XLM-RoBERTa or mBERT.

Social Network Analysis: Explore the network of interactions among users on social media platforms and analyze sentiment patterns within these networks. This can provide insights into influential users, opinion leaders, and the spread of sentiments within social networks.

Sentiment-Based Election Prediction: Investigate the possibility of using sentiment analysis as a tool for predicting election outcomes. Develop predictive models that leverage sentiment data from social media to forecast election results or to understand the potential impact of sentiments on voting behavior.

Comparison with Traditional Polling: Compare the results of sentiment analysis with traditional polling methods to assess the effectiveness and reliability of sentiment analysis as a complementary approach for gauging public sentiment. This could involve conducting surveys or interviews to validate the sentiment analysis findings.

References:

[1] R Tekchandani, R Joshi, “Sentiment Analysis of Twitter Data Using Machine Learning Techniques” , TIET Digital Repository.

[2] P. Garg, “Sentiment Analysis of Twitter Data using NLTK in Python”, Thapar University, June 2016.

[3] B. Pooja, K. Katyayini, G. Prasad et al, “Sentiment and Emotional Analysis of User Opinions on Twitter Data Using Machine Learning Techniques”, Anil Neerukonda Institute of Technology & Sciences.

[4] M. Lamberti, “Project Report Twitter Emotion Analysis”, Hong Kong University of Science and Technology.

[5] D Manu, “Lok Sabha Elections 2019: Analyzing the Online Political Battlefield in India” , Research in Computer Science and Engineering IIIT-Hyderabad, October 2020.

[6] SAURABH SHAHANE (2021), “Twitter Sentiments Dataset”, https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset

[7] ADRIT PAL, “Dataset of Indian Politics Tweets & Reactions”, https://www.kaggle.com/datasets/adritpal08/dataset-of-indian-politics-tweets-and-reactions

[8] Katta, P., Hegde, N.P.: A Hybrid Adaptive Neuro-Fuzzy Interface and Support Vector Machine Based Sentiment Analysis on Political Twitter Data. Int. J. Intell. Eng. Syst. 12, (2019). https://doi.org/10.22266/ijies2019.0228.17.

[9] Sharma, P., Moh, T.S.: Prediction of Indian election using sentiment analysis on Hindi Twitter. Proc. - 2016 IEEE Int. Conf. Big Data, Big Data 2016. 1966–1971 (2016). https://doi.org/10.1109/BIGDATA.2016.7840818.

[10] Glassman, M.E., Straus, J.R.: CRS Report for Congress Social Networking and Constituent Communication: Member Use of Twitter During a Two-Week Period in the 111 th Congress. (2009).

[11] Kim, S.-M., Hovy, E.: Determining the sentiment of opinions. 1367-es (2004). https://doi.org/10.3115/1220355.1220555.